高次元小標本におけるデータ解析の数理統計学的基礎とその応用

著者	青嶋 誠
発行年	2010
その他のタイトル	MATHEMATICAL STATISTICS FOR DATA ANALYSIS IN HIGH DIMENSION, LOW SAMPLE SIZE CONTEXT AND ITS APPLICATIONS
URL	http://hdl.handle.net/2241/107636

様式 C-19

科学研究費補助金研究成果報告書

平成２２年　３月３１日現在

研究種目：基盤研究（Ｂ）
研究期間：2006～2009
課題番号：１８３００９２
研究課題名（和文）　高次元小標本におけるデータ解析の数理統計学的基礎とその応用
研究課題名（英文）　MATHEMATICAL STATISTICS FOR DATA ANALYSIS IN HIGH DIMENSION, LOW SAMPLE SIZE CONTEXT AND ITS APPLICATIONS

研究代表者
青嶋　誠（AOSHIMA MAKOTO）
筑波大学・大学院数理物質科学研究科・教授
研究者番号：９０２４６６７９

研究成果の概要（和文）：　高次元小標本における高次元漸近理論を、非正規の一般的な設定のもとで構築した。高次元小標本データ特有の幾何学的な構造を発見した。従来型の PCA が高次元小標本で不一致性を引き起こすことを証明した。クロスデータ行列法とノイズ掃き出し法を提唱し、次元推定・固有値・漸近分布・固有ベクトル・主成分スコアの推定に、一致性をもつ解を与えた。クラスター分析と判別分析への応用を考え、前立腺がんのマイクロアレイデータの解析に役立てた。

研究成果の概要（英文）：We developed the high-dimension asymptotic theory for High Dimension, Low Sample Size (HDLSS) datasets under a general setup such as non-Gaussian distributions. We found several geometric structures of HDLSS datasets. We showed that the naïve PCA is inconsistent in the HDLSS context. We proposed effective inference methods called (1) the noise-reduction methodology, and (2) the cross-data-matrix methodology. By using those methodologies, we gave consistent estimation for intrinsic dimensionality, eigenvalues, their limiting distributions, PC directions and PC scores in the HDLSS context. We applied those methodologies to the discriminant analysis and the cluster analysis in HDLSS data situations from a microarray study of prostate cancer.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	3,300,000	990,000	4,290,000
２００７年度	2,800,000	840,000	3,640,000
２００８年度	4,300,000	1,290,000	5,590,000
２００９年度	4,300,000	1,290,000	5,590,000
年度			
総　計	14,700,000	4,410,000	19,110,000

研究分野：高次元統計解析
科研費の分科・細目：情報学・統計科学
キーワード：多変量解析、機械学習、パターン認識、モデル選択、ノイズ、生体生命情報学、マイクロアレイ、高次元データ

１．研究開始当初の背景

マイクロアレイデータなどの近代科学で扱うデータの一つの特徴は、データの次元数が標本数よりも遥かに大きな高次元小標本にある。高次元小標本においては大標本漸近理論に基づく多変量解析の理論は破綻する。高次元小標本データ空間には、無視できないほど大きなノイズが混在する。ノイズの影響を考えずに、大標本漸近理論による従来型の多変量解析を用いると、推測に誤った結果が導かれることは、実験的に報告されていた。しかしながら、高次元小標本における理論の整備が未開拓であったために、理論の開拓と新たな方法論の開発が急務とされていた。

２．研究の目的

高次元小標本データに対する漸近理論を新たに構築して、高次元小標本に特有な各種推測問題の解決を目指し、次の３つのテーマを目的に掲げた。

(1) 高次元小標本データ空間の幾何学的な解釈と標本数を伴う高次元小標本漸近理論の構築。
(2) 高次元空間に内在する潜在空間の次元の推定と、高次元小標本における主成分分析（PCA）の開発。
(3) 高次元小標本におけるパターン認識のための各種方法論の開発。

３．研究の方法

(1) 高次元空間にプロットされた正規分布の小標本は球面集中現象をもつ。これは次元の呪いの一種として知られ、Hall et al. (2005), Ahn et al. (2007) 等によって緩い非正規性のもとで数理統計学的な解釈が与えられようとしていた。研究目的(1)では、先行研究の致命的誤りを正すことから研究を始めた。高次元小標本データ空間に特有な幾何学的構造を理論的に解明するためには、標本数を伴う高次元小標本漸近理論を精密に構築する必要があった。研究代表者は、研究分担者で統計的漸近理論の世界的権威である赤平と研究を進め、ランダム行列の専門家である研究分担者の南とも研究打合せを行った。また、幾何学を専門とする研究分担者の田崎、川村とも定期的にセミナーを開き、統計学・確率論・幾何学を融合させ、最先端の理論と方法論を取り入れて研究を行った。
(2) 高次元空間に内在する潜在空間は巨大なノイズに埋もれている。本研究課題では、これを小標本で浮き彫りにする必要がある。従来型の PCA などの方法論は、巨大なノイズがもたらす次元の呪いによって不適切な解が出力されることが、先行研究の実験から報告されていた。研究目的(2)は、従来の方法論がなぜ次元の呪いを被るのかを数学的に解明することから研究を開始した。そのためには、高次元小標本データ空間のノイズの大きさと挙動を精密に解析する必要がある。研究代表者は研究分担者の赤平と連絡を密に取りながら、(1)で構築した高次元小標本漸近理論を駆使して、ノイズの見積りを精密に行い、従来型の方法論が被る次元の呪いの理論的な解明に努めた。高次元小標本に新たな方法論を開発するためには、内在する潜在空間の次元の推定と、固有値の推定が鍵となる。これには、逐次的なオンライン学習が必要になる。研究代表者は研究分担者で逐次推定論の専門家である小池と研究打合せを進めて、高次元小標本における新たな PCA の開発に努めた。
(3) 高次元小標本におけるパターン認識の方法論も、数理統計学の立場からの理論的な確立がなされているわけではなかった。従来の多変量解析にある判別分析やクラスター分析に若干の修正を加えたものや、理論的な保証がない方法論を、シミュレーション実験で比較するという先行研究が殆どであった。研究目的(3)では、あくまで理論的な保証のもとで、高次元小標本におけるパターン認識の方法論の構築を目指した。そのため、研究代表者は研究分担者の赤平と連絡を取って、高次元小標本漸近理論を駆使した方法論の開発に努めた。また、実用的な方法論を構築するために、研究分担者で医学を専門とする高橋と連絡を取り、マイクロアレイデータの解析に役立つ方法論の開発に努めた。非正則な場合のクラスタリングを扱うことになるため、研究代表者は研究分担者で非正則下の情報量の専門家である大谷内と連絡を取り、大標本漸近理論に基づく従来型の情報量に替わる新しい情報量を導入して、モデル選択のための新しい基準の構築に努めた。

４．研究成果

(1) 高次元小標本において、非正規な一般的設定のもと、大標本漸近理論に替わる新たな漸近理論として、標本数を伴う高次元小標本漸近理論を構築した。この理論を用いることで、先行研究では見出すことができなかった、高次元小標本データに特有な幾何学的構造を発見した。高次元小標本において、従来の多変量解析による推測の解が、次元の呪いをいかに被るかを理論的に解明し、推測の不一致性を数学的に証明した。さらには、従来の推定に標本数決定に関するある修正を

加えると、推測に一致性を回復することも数学的に証明した。例えば、高次元小標本データ空間の潜在空間における固有ベクトルの推定は、従来の多変量解析では９０度方向性を誤り一意に方向が定まらない、ということを証明できる。これに、本研究で与えた標本数決定に関する修正を施せば、一致性をもつ推定を与えることができる。しかし、従来型の多変量解析を修正するだけの方法論で得られる推定は、多くの場合、もはや、高次元小標本の枠組に存在しない。高次元小標本において一致性をもつ推定を得るためには、高次元小標本データに特有の新たな方法論の開発が必要になる。その鍵を握るのが、本研究で構築をした高次元小標本漸近理論であり、その意味で、本研究課題に関連する国内外の研究機関に与えるインパクトは大きい。

(2) 巨大なノイズに埋もれた高次元小標本データ空間の潜在空間を浮き彫りにするために、高次元小標本データに特有な新たな推測の方法論を、(1)で構築した高次元小標本漸近理論に基づいて開発した。推測に新しい２つの方法論を提唱した。一つは、「ノイズ掃き出し法」と名付けた。これは、高次元小標本データ空間のノイズの大きさと挙動を、(1)で構築した高次元小標本漸近理論を用いて精密に解析して、ノイズを掃き出して推測を行うという方法論である。いま一つは、「クロスデータ行列法」と名付けた。これは、データ行列を適当に分割して、これらをクロスで掛け合わせ、生成されるクロスデータ行列の特異値分解に基づくノンパラメトリックな方法論である。これら２つの方法論に基づく高次元小標本における推測が、一致性をもつことを、潜在空間の次元数・固有値・固有値の漸近分布・固有ベクトル・主成分得点について、理論的に証明した。本研究で提唱した２つの方法論は、ともに計算効率が非常に優れ、また推定の精度が大変に高く十分に実用性があるため、高次元データ解析を扱う近代科学の様々な応用分野に、今後、インパクトを与えていくと予想される。

(3) 高次元小標本におけるパターン認識として、(2)で提唱したノイズ掃き出し法に基づく判別分析と、クロスデータ行列法に基づくクラスター分析を提案した。それらの方法論の性能を理論的に証明し、計算時間においても極めて実用的であることを実験によって検証した。提案するクラスタリング手法を前立腺がんのマイクロアレイデータに適用して、腫瘍患者と正常患者を非常に高い正答率で分類することに成功した。また、高次元小標本においてモデル選択を行うための各種情報量を考えた。大標本漸近理論に裏打ちされた最尤法の枠組を外して多様なモデルを評価するために、カルバック・ライブラー情報量に替わる U－ダイバージェンスを考え、異常値に対する頑健性を考慮しβ-ダイバージェンスを扱った。最大β尤度推定に基づく非正則な場合のモデル構築を考え、予測に関するバイアス補正を行って、新しいモデル選択の基準を提案した。混合正規分布モデルを考え、最大β尤度推定のアルゴリズムを開発し、それに基づくクラスタリングを提案した。AIC、TIC と比較して性能を実験で検証した。提案したクラスタリング手法の性能が大変に高いため、今後、生命科学の分野の発展に、役立つことができれば望外の幸せである。

なお、本研究課題の研究成果を広く社会に発信するために、研究期間の２年目、３年目、４年目（最終年度）に、研究課題に関連するシンポジウムを開催した。毎回５０名程度の参加者に恵まれ、活発な研究発表と有意義な情報交換を行うことができる場となった。

① 日本学術振興会科学研究費による研究集会「統計的データ解析手法の評価と開発」, 広島大学, 2008年1月16-18日
② 日本学術振興会科学研究費による研究集会「Recent Advances in Statistical Inference-In Honor of Prof. Akahira」, 筑波大学, 2008年12月15-17日
③ 日本学術振興会科学研究費による研究集会「高次元データの統計学-理論・方法論・関連分野への応用-」, 筑波大学, 2009年12月14-16日

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３１件）

① Aoshima, M., Yata, K. (2010). Asymptotic second-order consistency for two-stage estimation methodologies and its applications. *Ann. Inst. Statist. Math.*, 査読有, 62, in press
② Yata, K., Aoshima, M. (2010). Intrinsic dimensionality estimation of high dimension, low sample size data with d-asymptotics. *Commun. Statist. -Theory and Meth.*, 査読有, 39, in press
③ Koike, K. (2010). Sequential estimation procedures for end points of support in a non-regular distribution. *Commun. Statist.*

-Theory and Meth., 査読有, 39, in press

④ Yata, K., Aoshima, M. (2009). PCA consistency for non-Gaussian data in high dimension, low sample size context. *Commun. Statist.-Theory and Meth.*, 査読有, 38, 2634-2652

⑤ 赤平昌文 (2008). 非正則推定における情報量の概念とその役割. 日本統計学会誌, 査読有, 37, 329-342

⑥ Akahira, M., Ohyauchi, N. (2007). A Bayesian view of the Hammersley-Chapman-Robbins-type inequality. *Statistics*, 査読有, 41, 137-144

〔学会発表〕(計３５件)

① Aoshima, M. Eigenvalue estimation in HDLSS context and its applications. 日本学術振興会日露共同研究プロジェクト研究集会, 広島大学, 2008年11月4日

② Aoshima,M. Intrinsic dimensionality estimation of high dimension, low sample size data with geometric representation. International IISA Conference, University of Connecticut, Connecticut, U.S.A., May 24, 2008

③ 大谷内奈穂. Information inequality bounds in non-regular estimation. 日本数学会秋季総合分科会, 統計数学分科会 特別講演, 東北大学, 2007 年 9 月 23 日

④ Koike, K. Sequential estimation of a location parameter for the location-scale family of distributions in non-regular case. The 56th Session of the International Statistical Institute, Lisboa Congress Centre, Lisboa, Portugal, August 27, 2007

⑤ Aoshima, M. Asymptotic second-order consistency for two-stage estimation methodologies and its applications. The 56th Session of the International Statistical Institute, Lisboa Congress Centre, Lisboa, Portugal, August 25, 2007

⑥ Aoshima, M. Asymptotic second-order consistency for two-stage methodologies via covariance structures. First International Workshop in Sequential Methodologies, Auburn, U.S.A., July 25, 2007

⑦ Akahira, M. The second order large-deviation efficiency for an exponential family of distributions. 日本学術振興会日露共同プロジェクト研究集会, 広島大学, 2006 年 8 月 7 日

〔図書〕(計３件)

① 青嶋　誠、他、電子情報通信学会、知識ベース：電子情報通信基礎、2010、印刷中

〔その他〕
ホームページ等
つくばリポジトリ：
https://www.tulips.tsukuba.ac.jp/portal/tulips-r.php

６．研究組織

(1)研究代表者
青嶋　誠 (AOSHIMA MAKOTO)
筑波大学･大学院数理物質科学研究科・教授
研究者番号：90246679

(2)研究分担者
赤平　昌文 (AKAHIRA MASAFUMI)
筑波大学・副学長
研究者番号：70017424
小池　健一 (KOIKE KEN-ICHI)
筑波大学・大学院数理物質科学研究科・准教授
研究者番号：90260471
大谷内　奈穂 (OHYAUCHI NAO)
筑波大学･大学院数理物質科学研究科･助教
研究者番号：40375374
田崎　博之 (TASAKI HIROYUKI)
筑波大学・大学院数理物質科学研究科・准教授
研究者番号：30179684
(H20→H21：連携研究者)
川村　一宏 (KAWAMURA KAZUHIRO)
筑波大学・大学院数理物質科学研究科・准教授
研究者番号：40204771
(H20→H21：連携研究者)
高橋　秀人 (TAKAHASHI HIDETO)
筑波大学・大学院人間総合科学研究科・准教授
研究者番号：80261808
(H20→H21：連携研究者)
南　就将 (MINAMI NARIYUKI)
慶應義塾大学・医学部・教授
研究者番号：10183964
(H20→H21：連携研究者)

(3)研究協力者
矢田　和善 (YATA KAZUYOSHI)
筑波大学・大学院数理物質科学研究科・博士後期課程３年